# The Quality of Local District Assessments Used in Nebraska's School-Based Teacher-Led Assessment and Reporting System (STARS)

Susan M. Brookhart, *Duquesne University*

*A sample of 293 local district assessments used in the Nebraska STARS (School-based Teacher-led Assessment and Reporting System), 147 from 2004 district mathematics assessment portfolios and 146 from 2003 reading assessment portfolios, was scored with a rubric evaluating their quality. Scorers were Nebraska educators with background and training in assessment. Raters reached an agreement criterion during a training session; however, analysis of a set of 30 assessments double-scored during the main scoring session indicated that the math ratings remained reliable during scoring, while the reading ratings did not. Therefore, this article presents results for the 147 mathematics assessments only. The quality of local mathematics assessments used in the Nebraska STARS was good overall. The majority were of high quality on characteristics that go to validity (alignment with standards, clarity to students, appropriateness of content). Professional development for Nebraska teachers is recommended on aspects of assessment related to reliability (sufficiency of information and scoring procedures).*

**Keywords:** assessment, validity, reliability

The State of Nebraska uses an assessment system based on standards of achievement for grades 4, 8, and 12 in order to monitor the progress of students in its public schools. The Nebraska School-based Teacher-led Assessment and Reporting System (STARS) is a unique state accountability system. School districts identify how they will measure and report student performance on content standards. They may select norm-referenced tests, develop criterion-referenced assessments, or use classroom assessments to measure state or state-approved content standards (Roschewski, 2004, p. 10). Nebraska is known nationally for its position that local districts should be responsible for the assessment systems that monitor the progress of the students they teach (Roschewski, Gallagher, & Isernhagen, 2001).

District systems for reporting student achievement are reviewed for quality, thus maintaining high standards for local assessment practices as well as supporting student achievement. The quality of district assessment systems has been evaluated annually since 2001 (Buckendahl, Plake, & Impara, 2004). Districts submitted portfolios describing their assessments and quality control systems in reading in 2001 and 2003 and in mathematics in 2002 and 2004. Six quality criteria (Plake, Impara, & Buckendahl, 2004) have been used to provide technical quality ratings for district assessment systems.

This study is the first look at the quality of the local assessments used in those systems. Until now, districts' assessment systems have been evaluated (Buckendahl et al., 2004), and the quality of individual local assessments has been assumed. This study addressed that assumption with empirical evidence. In 2003 each Nebraska school district included, as an appendix to its reading assessment portfolio, local assessments covering three reading standards each at grades 4, 8, and 11. (Twelfth-grade content standards are tested at grade 11.) Similarly, in 2004 districts included local mathematics assessments in their math assessment portfolios. The Nebraska Department of Education (NDE) had randomly selected and randomly assigned content standards for the local assessment samples to be included in district portfolios. However, these local assessments

*Susan M. Brookhart is Coordinator of Assessment and Evaluation, School of Education, Duquesne University, 600 Forbes Avenue, Pittsburgh, PA 15282; susanbrookhart@bresnan.net. Her areas of specialization are assessment, test reliability, and professional development.*

themselves had not been rated in the district portfolio rating process. Rather, they were considered examples as corroborating evidence for judging the districts' assessment process and system.

The research questions for this study were the following:

1. Of what quality are the local assessments used in the Nebraska STARS program?
2. What proportion of them is of sufficient quality to accurately measure student performance?
3. Is the quality of local assessments related to the quality of the district assessment system?
4. If cases are found where quality is not deemed sufficient, what professional development and/or feedback to teachers might be required to raise the quality of assessment to an acceptable level?

## Method

### *Sample Assessments*

Nebraska categorizes school districts into six classes based on grade level and population of territory. A random sample of 300 assessments was selected from the 2003 and 2004 district assessment portfolios stratified by school district class (ESS/NDE, 2003), 50 assessments for each of the two subjects at three grade levels. The researcher compiled a list of randomly selected districts (or in some cases consortia of districts that submitted joint assessment portfolios) for each class of school district. The number of districts to be sampled in each class was proportional to the number of districts of that class in the state. The researcher then supplied this list of randomly selected portfolios to the Nebraska Department of Education whose assessment office staff pulled sample assessments from the selected districts' portfolios, removed identifying information, and photocopied them for use in a scoring session.

Nebraska school districts collaborate on assessment design, often with the assistance of their regional Educational Service Unit. For this reason, the district assessment portfolios contained many duplicate assessments. If an assessment drawn from a selected portfolio duplicated an assessment already pulled for the sample, it was not used. The sampling proceeded to the next randomly selected district portfolio in the appropriate class, so that the resulting sample included 300 unique assessments. The NDE staff member who supervised the drawing of the sample estimated that the final sample of 300 assessments represented about one third to one half of the unique assessments in district reading and mathematics portfolios.

From the 300 assessments, 30 were randomly selected (5 each for 2 subjects at 3 grade levels); for these 30, two copies were included with the 300 assessments to be scored, thus embedding a double scoring study in the procedure. Some of the assessments did not include enough information to be scorable (for example, some districts had sent in incomplete copies of the assessments). The final sample size was 293 assessments, 147 in math and 146 in reading.

### *Participants*

Thirteen Nebraska educators, identified and invited by NDE, participated in the scoring workshop. Eight of these were graduates of the University of Nebraska-Lincoln (UNL) assessment cohort program (Lukin, Bandalos, Eckhout, & Mickelson, 2004). All were experienced educators who had worked extensively with assessment. Most had participated in planning the assessment system for their district or consortium of districts and were very familiar with the STARS procedures. These educators would be using the assessment rubrics as leaders of professional development around the state in the coming year, so the scoring session functioned as training for them as well as a rating session.

The two-day scoring session was observed by the NDE Director of Statewide Assessment and three NDE curriculum coordinators (reading, math, and social studies) and by two UNL professors, including the coordinator of the STARS overall evaluation, of which this study is one part. The observers were present at the request of NDE, to keep the process open.

### *Instrument (Rubric) Development*

Rubrics were developed with a dual purpose in mind. First, the rubrics were to be useful for the study described here. Second, the rubrics were to be sufficiently user-friendly and in line with other Nebraska STARS materials to be useful in district professional development. The goal was to study the quality of local assessments currently used in STARS, and then to use both the results and the rubrics to improve the quality of local assessments even further.

*Content Validity.* Several programs of research have codified the elements of quality assessment into rubrics. The rubrics for this study were developed by analyzing and comparing several existing sets of rubrics for judging the quality of classroom assessments: the Alternative Assessment Checklist (NWREL, 1998), the Classroom Assessment Quality Rubrics (Arter & Busick, 2001), the CRESST Language Arts Assignment Rubric (Matsumura, Garnier, Pascal, & Valdes, 2002), and rubrics developed for assessing student teachers' assessments (McConney & Ayres, 1998).

The author used the characteristics of high-quality assessment identified in the literature review to draft rubrics relevant to the purpose of the local assessments in STARS: to measure levels of achievement of state content standards. Current conceptions of reliability and validity for classroom assessments (Brookhart, 2003; Linn, Baker, & Dunbar, 1991) were also considered in drafting the rubrics. There are at least two additional criteria important to the quality of local assessments that could not be measured for this study: amount of student involvement and integration with instruction (opportunity to learn/practice). Investigating these would have required information outside of the scope of this study.

The first draft of the rubrics did not look like other STARS material. The NDE Director of Statewide Assessment suggested revisions that would make the rubrics consistent with other NDE assessment communications (NDE, 2003). The resulting rubrics had five traits: Alignment, Sufficiency, Clarity, Appropriateness, and Scoring Procedures. A copy of the rubrics used in the study is presented as Figure 1.

Both the text of the rubrics in Figure 1 and their operational definitions as used by the workshop participants were intended to be consistent with the STARS definitions of characteristics of assessment quality already in use for the reviews of district assessment portfolios. A recent NDE *STARS Update* (No. 14: NDE, 2004) had included information for districts about these characteristics for use in preparing their 2004 assessment portfolios. The participants were familiar with

*Nebraska Department of Education*
*July 13-14, 2004*

**ASSESSMENT CODE** _____ **DISTRICT** _____ **STANDARD** _____

**RATER** _____

*Circle the specific part of the criterion that is present and assign points for each criterion.*

| CRITERIA | 1-2 Points | 3-4 Points | 5-6 Points | SUMMARY POINTS |
|---|---|---|---|---|
| **ALIGNMENT** | Few of the assessment items/tasks reflect a match to the appropriate standard. | Some of the assessment items/tasks reflect a match to the appropriate standard(s). | Assessment items/tasks reflect a match to the appropriate standard(s). | |
| **SUFFICIENCY** | The essence of the standard is not represented, is ignored, or is poorly sampled. | The essence of the standard is represented by an inadequate number of score points and only at certain performance levels. | The essence of the standard is represented by an appropriate number of score points at all four performance levels (required by 2006-07). | |
| **CLARITY** | Few of the assessment tasks and/or directions are clear, complete, or unambiguous. | Some of the assessment tasks or the directions are clear, complete, or unambiguous. | The assessment tasks and the directions are clear, complete, and unambiguous. | |
| **APPROPRIATENESS** | Few of the assessment tasks are fair, at the appropriate cognitive level, or of an appropriate length. | Some of the assessment tasks are fair, at the appropriate cognitive level, or of an appropriate length. | Assessment tasks are fair, at the appropriate cognitive level, and of an appropriate length. | |
| **SCORING PROCEDURES** | Scoring procedures/rubrics are inadequate or not provided. | Scoring procedures/ rubrics are provided, but require judgments that would be hard to make fairly and consistently. | Scoring procedures/ rubrics are consistent with the assessment task and can be applied clearly and consistently. | |

FIGURE 1. Classroom assessment rating rubric.

this material from work in their own districts. Participants voiced appreciation about having a rubric that applied to individual assessments the same concepts with which they had been working at the system level. A description of the operational definitions of the five traits follows.

The Alignment rubric described alignment as reflecting a match to appropriate standards. This meant that the content of the test was an accurate reflection of the standard it claimed to measure. Copies of the Nebraska Reading/Writing and Mathematics Standards (NDE, 2000, 2001) were provided to each participant. The Nebraska Standards include example indicators that illustrate what kinds of content each standard is intended to include. Participants were instructed to read the relevant standard before rating each assessment even if they thought they were familiar with it.

Sufficiency was defined to mean an appropriate number of score points at all four performance levels: Beginning, Progressing, Proficient, and Advanced. The phrase "(required by 2006-07)" was inserted in the rubric to remind participants that the current review of district assessment systems only required sufficiency of information by standard, usually judged by considering the number of items or points per standard. To improve the quality of information, the state established the goal that by the 2006–2007 school year, district assessments would pay attention to the amount of information available at all levels. For example, tests should include some items that even Beginners would get correct and some items that only Advanced students would get correct, in order to accurately categorize students into four performance levels. *STARS Update 14* (NDE, 2004, p. 26) defined this to mean a minimum of 3

items (or 3 points) at each level for each standard (total 12 score points per standard), with more preferred, if the information was to be used for group information with low stakes to students. This definition of "sufficiency" was required for full points on this rubric.

Clarity was defined as clarity to the student. The assessment tasks and directions were read to see if they were clear and complete from the student's point of view. This was operationalized informally in the workshop by directing participants to ask themselves, "Would the kid know what to do?"

The Appropriateness rubric subsumed two kinds of appropriateness. Fairness meant that diverse learners would find the assessment's items or tasks to be appropriate for them: not offensive and within their cultural experience. "Cognitive level" and length were attempts to describe assessments pitched at the appropriate developmental level; however, using the word "developmental" had been ruled out because of some other connotations in the STARS system. For example, a fourth-grade test should be appropriate for fourth-grade students.

The Scoring Procedures rubric asked participants to check that scoring procedures were consistent with the assessment task and could be applied clearly and consistently. This meant first checking to see whether scoring procedures were indeed supplied with the assessment. If they were, the next step was to see if the scoring procedures were in line with the standard they were intended to measure and were clear enough to be reliably applied.

For use in the scoring session, the rubrics were printed on one page. Space was provided for recording codes for the assessment ID number, district, standard, and rater. The resulting one-page instrument (Figure 1) was filled in for four example assessments, to demonstrate how to use the form as well as to describe the example assessments. One sheet per rater per assessment was used for scoring training assessments and for the main study rating.

*Reliability.* Three levels of performance were described for each trait: low (1–2 points), medium (3–4 points), and high (5–6 points). Full points (2, 4, or 6) were awarded if the criteria for that level were met in full, the lesser point amount (1, 3, or 5) if the description applied but was not met in full. The first day of the scoring session was a training day. First, four example assessments were discussed, and then five training assessments were scored independently. Rater agreement overall (13 raters rating 5 training assessments on 5 criteria) was 86% exact agreement on category (low/medium/high) and 66% exact agreement on point value (1–6). Agreement on category was not the same as exact plus adjacent agreement. If adjacent agreement was within category, it counted as agreement (e.g., 5 and 6 were an agreement); however, if adjacent agreement was not within category, it did not (e.g., 4 and 5 were not an agreement).

The training was planned to take a half of the first day (3 of the 6 hours) of the workshop. However, the discussion of both example and training assessments focused on clarifying the criteria and seemed to represent growing participant understanding of both the rubrics and their task. Therefore, the researcher did not cut off discussion in order to speed up the training. The training session took almost all of the first day. Raters scored the remaining assessments independently on the second day.

After the scoring session, additional reliability analyses were possible from the set of assessments that had been scored twice. Results from the 30 double-scored assessments indicated that after the initial training to criterion, math ratings remained more reliable than reading ratings. For the 30 double-scored assessments, exact agreement on category (low/medium/high) was 73% for math and 60% for reading. Exact agreement on point value (1–6) was 48% for Math and 37% for Reading. Intraclass correlations for total score (for a single rater) were .59 for math and −.13 for reading. Where subsets of double-scored assessments were scored by the same two raters (3 math subsets and 2 reading subsets, a total of 20 of the 30 assessments), generalizability analyses were possible (Crick & Brennan, 2001). Generalizability values for one rater ranged from .78 to .90 (median = .81) for relative decisions and .65 to .88 (median = .81) for absolute decisions in math and were .00 for reading. Therefore, this article presents only the results describing the quality of math assessments.

## Results

*Research Question 1: Of What Quality Are the Local Assessments Used in the Nebraska STARS Program?*

The quality of local mathematics assessments used in the Nebraska STARS was good overall. Table 1 presents the means and standard deviations by grade. For three of the criteria (Alignment, mean = 5.48, *SD* = .94; Clarity, mean = 5.62, *SD* = .72; and Appropriateness, mean = 5.38, *SD* = .89), the mean rating was in the top rubric category. Ratings for Scoring Procedures were lower (mean = 2.87, *SD* = 1.65), mostly because scoring procedures were not provided for many of the assessments. One of the participants described this as an "easy fix," and pointed out that the rubrics would be an effective way of communicating with districts the importance of including scoring procedures in local assessments. Scores for Sufficiency (mean = 4.73, *SD* = .91) also indicated an area for improvement. Many of the assessments did not have enough information available at all performance levels (Beginning, Progressing, Proficient, and Advanced) to reliably classify students at all performance categories.

*Research Question 2. What Proportion of Assessments Is of Sufficient Quality to Accurately Measure Student Performance?*

Tables 2 through 6 present the percentage of all math assessments at each level for Alignment, Sufficiency, Clarity, Appropriateness, and Scoring Procedures. Results disaggregated by grade level showed similar patterns. Overall, approximately 82% of math assessments were comprised of items or tasks reflecting a match to the standard they were intended to measure. Approximately 59% had an appropriate number of score points at all four performance levels (Beginning, Progressing, Proficient, and Advanced). Approximately 88% of assessments had tasks and directions that were clear to students. Approximately 82% of assessments were judged appropriate (fair for all students, and of an appropriate level and length for the intended grade).

Approximately 21% specified clear scoring procedures consistent with the task; 70% had scoring procedures that were either unclear or not provided.

**Table 1. Mean Mathematics Assessment Ratings**

| Grade | | Alignment | Sufficiency | Clarity | Appropriateness | Scoring Procedures |
|---|---|---|---|---|---|---|
| 4 | Mean | 5.67 | 4.96 | 5.59 | 5.43 | 3.14 |
| | *SD* | .69 | .84 | .73 | .76 | 1.79 |
| | n | 49 | 49 | 49 | 49 | 49 |
| 8 | Mean | 5.42 | 4.58 | 5.84 | 5.24 | 3.08 |
| | *SD* | .88 | .81 | .43 | .92 | 1.72 |
| | n | 50 | 50 | 49 | 50 | 50 |
| 11 | Mean | 5.33 | 4.65 | 5.44 | 5.48 | 2.37 |
| | *SD* | 1.17 | 1.04 | .87 | .97 | 1.32 |
| | n | 48 | 48 | 48 | 48 | 48 |
| Total | Mean | 5.48 | 4.73 | 5.62 | 5.38 | 2.87 |
| | *SD* | .94 | .91 | .72 | .89 | 1.65 |
| | n | 147 | 147 | 146 | 147 | 147 |

**Table 2. Alignment—Number and Percentage of Nebraska Mathematics Assessments at Each Level of Quality (Total = 147 Assessments)**

| Quality-Level Description | Points | Number of Assessments | Percentage of Assessments |
|---|---|---|---|
| Assessment items/tasks reflect a match to the appropriate standard(s). | 6<br>5 | 105<br>15 | 71.4<br>10.2 |
| Some of the assessment items/ tasks reflect a match to the appropriate standard(s). | 4<br>3 | 22<br>2 | 15.0<br>1.4 |
| Few of the assessment items/ tasks reflect a match to the appropriate standard. | 2<br>1 | 3<br>0 | 2.0<br>0.0 |

**Table 3. Sufficiency—Number and Percentage of Nebraska Mathematics Assessments at Each Level of Quality (Total = 147 Assessments)**

| Quality-Level Description | Points | Number of Assessments | Percentage of Assessments |
|---|---|---|---|
| The essence of the standard is represented by an appropriate number of score points at all four performance levels (required by 2006–2007). | 6<br>5 | 31<br>56 | 21.1<br>38.1 |
| The essence of the standard is represented by an inadequate number of score points and only at certain performance levels. | 4<br>3 | 53<br>3 | 36.1<br>2.0 |
| The essence of the standard is not represented, is ignored, or is poorly sampled. | 2<br>1 | 4<br>0 | 2.7<br>0.0 |

This figure may underestimate the quality of the local assessments' scoring procedures. It is possible that some districts did not provide scoring procedures when the state asked for assessments to be included in their portfolios because they interpreted that request to mean providing a copy of just the student instrument. Thus answer keys or other scoring procedures that were not available for rating may have existed for some assessments.

In summary, the majority of local assessments were of sufficient quality on characteristics that go to validity (alignment with standards, clarity to students, appropriateness of content). More work needs to be done to raise the quality of local assessments on aspects related to reliability (sufficiency of information and scoring procedures).

*Research Question 3: Is the Quality of Local Assessments Related to the Quality of the District Assessment System?*

The overall quality of assessment systems was rated separately for each district at grades 4, 8, and 11 with a procedure described by Buckendahl et al. (2004). The state provided these overall ratings to the author of this study. These ratings made it possible to investigate whether the quality of local assessments was related to the quality of the assessment system in a district. In this study of local assessment quality, most districts were represented by two or three sampled assessments. The median scores for assessments in each district at each grade were matched with that district/grade's assessment system rating (1–5, unacceptable through exemplary, based on the 2004 math portfolio ratings). Spearman correlations were calculated. Table 7 presents the results.

None of the relationships reached significance after a Bonferroni adjustment for doing five tests was applied. There are many possible reasons for the lack of observed relationship between the quality of the district's assessment system, as reflected in the portfolio rating, and the quality of the local assessments. Lack of variability was clearly one of the reasons. In math,

**Table 4. Clarity—Number and Percentage of Nebraska Mathematics Assessments at Each Level of Quality (Total = 146 Assessments)**

| Quality-Level Description | Points | Number of Assessments | Percentage of Assessments |
|---|---|---|---|
| The assessment tasks and the directions are clear, complete, and unambiguous. | 6 | 110 | 75.3 |
| | 5 | 18 | 12.3 |
| Some of the assessment tasks or the directions are clear, complete, or unambiguous. | 4 | 17 | 11.6 |
| | 3 | 1 | 0.7 |
| Few of the assessment tasks and/or directions are clear, complete, or unambiguous. | 2 | 0 | 0.0 |
| | 1 | 0 | 0.0 |

**Table 5. Appropriateness—Number and Percentage of Nebraska Mathematics Assessments at Each Level of Quality (Total = 147 Assessments)**

| Quality-Level Description | Points | Number of Assessments | Percentage of Assessments |
|---|---|---|---|
| Assessment tasks are fair, at the appropriate cognitive level, and of an appropriate length. | 6 | 89 | 60.5 |
| | 5 | 31 | 21.1 |
| Some of the assessment tasks are fair, at the appropriate cognitive level, or of an appropriate length. | 4 | 22 | 15.0 |
| | 3 | 4 | 2.7 |
| Few of the assessment tasks are fair, at the appropriate cognitive level, or of an appropriate length. | 2 | 1 | 0.7 |
| | 1 | 0 | 0.0 |

**Table 6. Scoring Procedures—Number and Percentage of Nebraska Mathematics Assessments at Each Level of Quality (Total = 147 Assessments)**

| Quality-Level Description | Points | Number of Assessments | Percentage of Assessments |
|---|---|---|---|
| Scoring procedures/rubrics are consistent with the assessment task and can be applied clearly and consistently. | 6 | 26 | 17.7 |
| | 5 | 5 | 3.4 |
| Scoring procedures/rubrics are provided, but require judgments that would be be hard to make fairly and consistently. | 4 | 8 | 5.4 |
| | 3 | 5 | 3.4 |
| Scoring procedures/rubrics are inadequate or not provided. | 2 | 91 | 61.9 |
| | 1 | 12 | 8.2 |

96% of the district assessment system ratings were Very Good (16%) or Exemplary (80%). As Tables 2 through 6 showed, most of the ratings of individual assessments were also clustered at the top, although not quite as dramatically as the assessment system ratings, and the median ratings for the 51 district/grade samples formed similar distributions.

In addition to this statistical reason for lack of a relationship between district assessment systems and assessments, there are several other possible explanations. It may be that the district assessment system rating was too generalized a measure to relate to specific rubric scores; the overall district assessment system ratings were based on decision rules aggregating six criteria. It may be that the authors of the local assessments were equally capable, regardless of how well the district system to review and approve assessments functioned. It may be that portfolio ratings are partly a measure of how good the district is at expressing what it did to ensure assessment quality as well as a measure of the system quality itself. It is not possible to say which, if any, of these additional explanations might be true.

The correlation analysis has implications for professional development efforts (below). If the quality of local assessment instruments had been poorer in districts that needed help with their assessment system plans as well, that would have suggested a need for professional development in classroom assessment and in district assessment planning targeted to particular districts. The fact that the quality of the local assessment instruments was not related to the quality of the district assessment plans suggests a need for across-the-board professional development in classroom assessment.

*Research Question 4: Professional Development Needs*

The fourth question—If cases are found where quality is not deemed sufficient, what professional development and/or feedback to teachers might be required to raise the quality of assessment to an acceptable level?—was addressed in two ways. First, the rating patterns on the various rubric criteria gave a general indication of the kinds of improvements needed. Teachers were quite good at matching assessments to standards. This is not surprising, because selecting or writing assignments that match learning targets is something teachers do every time they write a lesson plan for student work. This study's results for Sufficiency and Scoring Procedures indicated that teachers would benefit from professional

**Table 7. Correlations beween District Assessment Portfolio Ratings and Median Ratings of Assessments Sampled from Those Portfolios**

| Assessment Quality Criterion | Math (2004) n = 51 |
|---|---|
| Alignment | −.26 |
| Sufficiency | .29 |
| Clarity | −.04 |
| Appropriateness | .12 |
| Scoring Procedures | .04 |

development on the ways in which scoring operates to turn student performance on that work into measurement of achievement.

Second, participants in the scoring session responded to questions about professional development in a debriefing. Professional development themes from these two sources included:

- Looking at sample assessments and discussing their quality was a powerful professional development tool.
- The rubric itself was valuable, making quality criteria explicit. The educators appreciated the guidance that the rubrics gave. The use of local assessments in the STARS program makes the quality of local district assessments a matter of importance to them.
- There was support and approval for state involvement in improving assessment literacy.

These educators described how having the criteria expressed in the rubrics and examples led to improved assessment literacy for them. There was enthusiastic support for the idea that discussions on example assessments would increase assessment literacy in districts, in a similar manner to the way they felt their assessment literacy increased with participation in this study.

Beyond raising the general level of assessment literacy among Nebraska educators, participants were asked to give specific suggestions for professional development work, based on the assessments they had seen. They called for professional development regarding what "sufficiency" is and how to tap performance with items or tasks at four levels. They observed that many of the assessments were actually mastery/nonmastery tests, with very little coverage at the beginning or advanced levels. They also suggested professional development about why scoring procedures are necessary, and how to write them well. Professional development about alignment or matching items or tasks to standard was also suggested. Teachers' interpretations of the standards need to "capture the essence" of the standard, not slavishly represent all the example indicators in the state standards. In the case of some assessments, participants felt they did not represent a deep knowledge base on the part of the assessment writers. Formatting was also suggested as a professional development topic. The design and presentation of questions, the size and type of answer spaces, and so on, communicate expectations to students. Participants also suggested that Nebraska teachers receive some guidance on what constitutes "appropriateness" (vocabulary, length, appropriateness of content to expected level of understanding and performance) over a range of kinds of standards. Some state content standards are broad, others are more narrow, and standards are written at different cognitive levels.

## Discussion

The Nebraska STARS has received national attention since its inception (Roschewski et al., 2001). Local assessment has a prominent role in state accountability reporting. While districts' assessment systems have been evaluated (Buckendahl et al., 2004), the quality of individual local assessments has been assumed. This study investigated that assumption. The majority of local mathematics assessments were of sufficient alignment, clarity, and appropriateness to warrant attention to their results. The results for sufficiency, coupled with participants' observations about lack of sufficiency being more of an issue for low and high performers, suggest that the quality of assessments may be of sufficient quality at the middle of the range of student performance. That is, Nebraska can be confident about the accuracy of estimation of the percentage of students above and below the Progressing/Proficient cutpoint; there is less confidence about the precise percentage of Beginning and Advanced students.

Even with a need for increased reliability of rubric use (similar to the development of reliability in the assessment system portfolio process; Buckendahl et al., 2004), this study has identified two areas that can with some confidence be shared with Nebraska school districts as areas for improvement. Attention should be given to developing explicit scoring procedures for local assessments and to sufficient coverage at all four performance levels: Beginning, Progressing, Proficient, and Advanced. Suggestions for professional development from participants included attention to the following: sufficiency, scoring procedures, alignment (especially interpretation of standards), depth of content knowledge, formatting, and appropriateness.

Using the rubrics could raise the quality of assessments in Nebraska even further, as teachers apply the rubrics in their local assessment processes. Additional work needs to be done to ensure that assessments in all subjects, not just mathematics, can be reliably judged with the rubrics. During the debriefing, participants made suggestions, not reported here, for some editorial changes that might make the rubrics clearer. Further discussion among Nebraska teachers about the meaning of the criteria would be helpful, too. Some of that discussion is beginning with a series of 10 "Assessing the Assessments" workshops being held around the state in 2004–2005.

Since this study was designed, another set of criteria and rubrics for judging alignment of state tests to standards, those used by Achieve, Inc., was published (Resnick, Rothman, Slattery, & Vranek, 2003-04). Achieve's rubrics would not have been suitable for this Nebraska study because the Achieve ratings included the possibility of judging the *standards* as not specific enough, and that was not in the purview of this study. Also, it assumed that multiple standards might be covered by a test, which is different from the way Nebraska districts sent in sample assessments by individual standard. However, it is important to note that the criteria used were congruent with the criteria used in this Nebraska study.

The five states in Achieve's sample were found to have items that mapped well for content and performance, but did not fare well on balance and range. In addition, the most challenging standards were poorly represented on tests. These results parallel the findings of

this study of Nebraska local assessments, where sufficiency of information at all performance levels was the most serious lack. Thus it seems that Nebraska has identified an area for improvement that it shares with other states around the country.

In summary, the quality of local district mathematics assessments used in Nebraska's STARS was generally good. Nebraska educators used a rubric developed to comport with other STARS materials to make judgments of a sample of reading and mathematics assessments. Reliability of ratings was a concern, especially for reading, and participants recommended ways in which the rubric might be revised to improve clarity. Because of reliability concerns, this article reported the results for mathematics assessments only.

Participants found the rating exercise itself to be good professional development. They were enthusiastic about using the rubric developed for this study in work with Nebraska teachers to raise assessment literacy and assessment quality statewide, especially around sufficiency and scoring issues. Participants felt that further work with this rubric would help improve local assessment quality. The quality of local district assessments is an essential part of an accountability system that relies on locally selected assessments rather than a common state assessment.

## References


Arter, J. A., & Busick, K. U. (2001). *Practice with student-involved classroom assessment*. Portland, OR: Assessment Training Institute.

Brookhart, S. M. (2003). Developing measurement theory for classroom assessment purposes and uses. *Educational Measurement: Issues and Practice, 22*(3), 5–12.

Buckendahl, C. W., Plake, B. S., & Impara, J. C. (2004). A strategy for evaluating district developed assessments for state accountability. *Educational Measurement: Issues and Practice, 23*(2), 17–25.

Crick, J. E., & Brennan, R. L. (2001, January). GENOVA (version 3.1). Available: http://www.uiowa.edu/~itp/pages/SWGENOVA.SHTML.

Education Support Services, Nebraska Department of Education (2003). Statistics and facts about Nebraska schools. Available: http://ess.nde.state.ne.us/DataCenter/PDF/0203/0203StatsFacts.pdf.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20,* 15–21.

Lukin, L. E., Bandalos, D. L., Eckhout, T. J., & Mickelson, K. (2004). Facilitating the development of assessment literacy. *Educational Measurement: Issues and Practice, 23*(2), 26–32.

Matsumura, L. C., Garnier, J., Pascal, J., & Valdes, R. (2002). Measuring instructional quality in accountability systems: Classroom assignments and student achievement. *Educational Assessment, 8,* 207–229.

McConney, A., & Ayres, R. R. (1998). Assessing student teachers' assessments. *Journal of Teacher Education, 49,* 140–150.

Nebraska Department of Education (2000). *Nebraska Mathematics Standards Grades K–12*. Available: http://www.nde.state.ne.us/ndestandards/documents/MathematicsStandards.pdf.

Nebraska Department of Education (2001). *Nebraska Reading/Writing Standards Grades K–12*. Available: http://www.nde.state.ne.us/ndestandards/documents/ReadingWritingStandards.pdf.

Nebraska Department of Education (2003). *The school district assessment portfolio instructions and suggestions*.

Nebraska Department of Education (2004). *STARS Update #14*. Available: http://www.nde.state.ne.us/stars/documents/Update14.pdf.

Northwest Regional Educational Laboratory. (1998). *Toolkit 98*. Portland, OR: Author.

Plake, B. S., Impara, J. C., & Buckendahl, C. W. (2004). Technical quality criteria for evaluating district assessment portfolios used in the Nebraska STARS. *Educational Measurement: Issues and Practice, 23*(2), 12–16.

Resnick, L. B., Rothman, R., Slattery, J. B., & Vranek, J. L. (2003–2004). Benchmarking and alignment of standards and testing. *Educational Assessment, 9,* 1–27.

Roschewski, P. (2004). History and background of Nebraska's School-based Teacher-led Assessment and Reporting System (STARS). *Educational Measurement: Issues and Practice, 23*(2), 9–11.

Roschewski, P., Gallagher, C., & Isernhagen, J. (2001). Nebraskans reach for the STARS. *Phi Delta Kappan, 82,* 611–615.